Journal of Information Science 2008; 34; 290 originally published online Jan 21, 2008;

DOI: 10.1177/0165551507084141

An online version of this article can be found at: http://jis.sagepub.com/cgi/content/abstract/34/3/290

C. Jenkins, C. Oppenheim, S.Probets and B. Hubbard

RoMEO Studies 7*: Creation of a Controlled Vocabulary to Analyse Copyright Transfer Agreements

Celia Jenkins; Charles Oppenheim; Steve Probets;

Department of Information Science, Loughborough University, Loughborough, Leics. LE11 3TU

Bill Hubbard

SHERPA, University of Nottingham, University Park, Nottingham NG7 2RD

***** Part 6, E. Gadd, C. Oppenheim and S. Probets, RoMEO studies 6: rights metadata in open archiving; *Program* 38(1) (2004) 5-14.

Abstract

This paper describes the process of creating a controlled vocabulary which can be used to systematically analyse the copyright transfer agreements (CTAs) of journal publishers with regard to self-archiving. The analysis formed the basis of the newly created Copyright Knowledge Bank of publishers' self-archiving policies. Self-archiving terms appearing in publishers' CTAs were identified and classified, with these then being simplified, merged, and discarded to form a definitive list. The controlled vocabulary consists of three categories that describe 'what' can be self-archived, the 'conditions' of self-archiving and the 'restrictions' of self-archiving. Condition terms include specifications such as 'where' an article can be self archived, restriction terms include specifications such as 'when' the article can be self archived. Additional information on any of these terms appears in 'free-text' fields. Although this controlled vocabulary provides an effective way of analysing CTAs, it will need to be continually reviewed and updated in light of any major new additions to the terms used in publishers' copyright and self-archiving policies.

1. Introduction

This paper discusses the process involved in developing a controlled vocabulary to be used to analyse journal publishers' self-archiving policies. Self-archiving terms and conditions were extracted from publisher's policies and used to form a controlled vocabulary. This was then used to systematically analyse publishers' copyright transfer agreements (CTAs), with this data then being input into a new Copyright Knowledge Bank (CKB) database. The CKB database is effectively a revised and updated version of the existing SHERPA/RoMEO database. The creation of this database is one of the deliverables of the JISC-SURF 'Partnering on Copyright' programme [1], which focused on copyright management with regards to open access and self-archiving.
In order to analyse the CTA data in a systematic and logical way, it was important to develop a controlled vocabulary through a mixture of simplifying, clarifying, merging and discarding terms used in actual CTAs. The resulting definitive, reduced list of terms made it easier to compare publishers' self-archiving policies and to develop a database that was consistent, accurate and clear. An XML schema describing the controlled vocabulary was also developed so that a publisher's CTA could be represented in a machine readable form if necessary.

2. Background

When an author submits a scholarly article for publishing in a journal, the author normally has to sign an agreement (the so-called Copyright Transfer Agreement, or CTA), which sets out the assignment of the copyright in the article to the publisher (or, in some cases, gives the publisher a licence to publish the article). It normally imposes certain terms and conditions on what the author can subsequently do with the published article in terms of self-archiving it in a digital repository. As a result of the JISC-funded RoMEO project [2], which took place August 2002 – September 2003, a list was compiled of over 70 worldwide journal publishers' CTAs along with details on whether they allowed the self-archiving of pre-prints, post-prints or both. This listing has become a well-known and heavily-used resource and has been developed and is now hosted and maintained by SHERPA [3]. The SHERPA/RoMEO database is used by authors and repository administrators when considering whether an article can be mounted on an institutional repository. However, it was felt that the SHERPA/RoMEO database of publishers' self-archiving policies, as it is now known, needed to be developed further so as to improve both its coverage and the level of information it provides on a specific CTAs . One path for development would be achieved by making the analysis of publishers' CTAs more comprehensive and structured, through the development and implementation of the controlled vocabulary.
Initially, this controlled vocabulary would be used to analyse the CTAs of publishers already in the SHERPA/RoMEO database, with the resulting data being input into the CKB. In future, the controlled vocabulary will be used to facilitate analysis of publishers' CTAs, as suggested by the self-archiving community, i.e., authors, librarians and publishers.

3. Developing the controlled vocabulary

A formal method for creating the controlled vocabulary was not employed, however the process taken did involve a number of iterations and refinements to ensure the terms used in the vocabulary were

wide ranging enough to cover all eventualities and to describe the CTAs in sufficient detail, yet concise enough so as to eliminate redundancy and provide a usable set of terms. The first stage in developing the controlled vocabulary was to become familiar with journal publishers' CTAs and their self-archiving policies. Through an initial analysis of a small number of CTAs, common sets of self-archiving terms were identified and categorised.

The initial analysis identified a number of terms used by publishers to describe their self-archiving policies. These were identified and grouped into categories. These categories were:

- *What*

  This included terms used in CTAs such as 'pre-prints', 'post-prints', and 'title and abstract/summary of the article'.

- *Where*

  This included terms such as 'author's personal Web site', 'employer's web homepage', and 'public repository', amongst others.

- *What version*

  This included terms such as 'author's own version, and 'publisher's formatted version'.

- *Requirements* – i.e., what the author has to do for self-archiving to be permitted by the publisher.

  This included requirements of the form: 'link to online abstract in journal or entry page of journal'; 'label the pre-print with the date and a statement that the paper is not yet published'; and 'copyright and citation notice to be embedded within full text file and in accompanying citation display' among many others.

This initial list of categories and common self-archiving terms formed a basis on which to build the controlled vocabulary.

The next step was to rigorously analyse the CTAs of all publishers appearing in the SHERPA/RoMEO database using the initial controlled vocabulary. SHERPA/RoMEO had already analysed CTAs and produced a set of standardised phrases to summarise different conditions and restrictions. The challenge was to allow greater granularity in the analysis and application of the controlled vocabulary than currently offered, while maintaining usability. This would enable more precise detail of conditions and restrictions outlined in CTAs to be presented to users of the database. Both authors and repository administrators are potential users of the database. It was therefore felt that presenting a more granular analysis of CTAs based on a controlled vocabulary would make it easier for e-prints to be managed in such as way that depositors and repository administrators could meet the terms outlined in a publisher's CTA. In addition, it was felt that a controlled vocabulary would be required to support a logical representation of CTAs that would be needed for any future machine to machine (M2M) interface to the database. An M2M interface could enable services embedded in institutional or subject-based repository software to automatically query the CTA database.

During the process it became clear that the self-archiving terms and definitions had to be simple yet unambiguous. It quickly became apparent that the creation of an effective controlled vocabulary would not be straightforward. There were two main reasons for this:

- Some publishers did not have their CTAs, or details of their self-archiving practices, available on-line, consquently some of the information had to come from paper copies of the publishing agreements and some from e-mails between the SHERPA project manager and the publishers in question. This introduced delays in obtaining information.
- Analysis of CTAs can be complex. Some of the CTAs are ambiguous or do not make the publisher's position on self-archiving clear or easy to understand.

A number of complexities were encountered during the development of the final version of the controlled vocabulary. The main challenges involved:

1. Reducing and defining the controlled vocabulary terms.

Publishers themselves use a variety of self-archiving-related terms, with very little standardisation or consistency. As a result, a large number of terms were used. It was, therefore, vital to produce a reduced definitive list of self-archiving terms to be used as the controlled vocabulary.

In addition to the number of different terms used, the majority of these were not defined by the publisher. In fact, some publishers seemed to use terms which were inappropriate and, under further investigation, did not appear to be what the publishers actually intended. Definitions for the terms had to be created that would make their meaning clear, not just for those wanting to discover the self-archiving policies of publishers, but also for those who will use the controlled vocabulary to analyse publishing agreements in future. The lack of clarity was particularly noticeable if CTAs specified 'where' authors may mount their work.

2. Aligning the controlled vocabulary with the technical considerations of the database.
As well as identifying categories and terms, it was necessary to look at how these may be aligned with the technical considerations of the database. It was decided early on in the project to base the CKB on the existing structure of the SHERPA/RoMEO database. This structure is flat rather than hierarchical, which makes it easier to work with and maintain. Building on the existing successful database would save on resources and help ensure that the CKB would stay manageable. It was therefore necessary to make sure that the controlled vocabulary could be represented without using an overly hierarchical structure.

4. Structure of the CKB

The current SHERPA/RoMEO database has a two-tier structure, segmenting a publisher's self-archiving policy into 'pre-print' and 'post-print' on the top level, and 'conditions' and 'restrictions' on a second level (i.e., within each of these categories). As the aim of the CKB project was to create a database which would provide more comprehensive information, a balance had to be made between, on the one hand, ensuring that the controlled vocabulary would satisfactorily describe self-archiving policies and, on the other, ensuring that it would fit within a structure similar to that of the SHERPA/RoMEO database.
The technical considerations therefore had an impact on the choice of terms, especially those concerning *what* could be self-archived. Originally, the terms describing what could be self-archived were separate from the terms describing what version could be self-archived. Versions were included as some CTAs specifically considered versions of the manuscript prior to submission to the publisher, after submission but before peer review, during the peer review process, etc. As a consequence of considering the need for a simple two-tier database structure, and to aid simplifying the presentation of the information to users of the database, a decision was made to merge the 'what version' into the 'what' category, with additional conditions indicating if there are any version-based conditions that need to be satisfied. Similarly, terms describing 'where' work could be self-archived were placed in a separate category, but later also became part of the 'conditions' category.

5. The controlled vocabulary

As a result of the CTA analysis and technical considerations of the CKB, it was decided that the controlled vocabulary would consist of three main categories:

- **'What'** can be self-archived, to incorporate 'what version'.
- What are the **'conditions'** of self-archiving, to include 'where' it can be self-archived.
- What are the '**restrictions'** to self-archiving, to include 'when' it can be self-archived.

Ensuring that the controlled vocabulary can represent CTAs in a simplified, easy-to-understand way is very important. However, as the CTAs can be complex documents with numerous elaborations and specifications, it was felt important that all of the information represented in the CTA should be available if desired. Therefore the controlled vocabulary was developed to allow for the inclusion of text

which gives more details about a publisher's self-archiving policy, through the creation of 'free-text fields'. These fields, which could perhaps be more accurately thought of as metadata rather than controlled vocabulary terms, contain text which either provides information which cannot be accurately represented using a controlled vocabulary term, or provides more detailed information regarding a particular term used. However as the CKB aims to provide details of all CTAs it was deemed necessary to include these free text fields.
The text found in the 'free-text fields' was, in the main, taken directly from CTAs and/or self-archiving policy documentation provided by the publisher, and includes both statements and links to the electronic version of the original CTA and/or an electronic document explaining the publisher's self-archiving policy where available. This information enables users of the CKB to easily find out further details by looking at the actual CTA or by contacting the publisher in question. It is recommended that users do seek further information if required as the CKB is only meant as a *guide* to publishers' self-archiving policies. The interface to the CKB was designed to reinforce this point.

Providing the actual words used by publishers also ensures that users of the CKB can make up their own minds about what is meant by the terms used by publishers. It would also allow for particular terms to be easily reinstated into the logic of the CKB. For example, if a publisher provides a clear explanation of a term which is subsequently adopted by several publishers, this could be added to the controlled vocabulary.
There are five free-text fields used in the CKB, each containing information relating to specific categories.
The five free-text fields are:

- 'what & where'
- 'specific statement'
- 'conditions'
- 'restrictions'
- 'additional information'

In the following sections of this paper, the categories and self-archiving terms of the final version of the controlled vocabulary are examined.

5.1. What can be self-archived?

In compliance with the two-level structure of the CKB, the top level consists of the 'what' category, which comprises of three 'types' of work which can be self-archived, these being:

- Type 1. Pre-print

  This was defined to be the primary, draft version of the Work, up to and during peer-review.
- Type 2 Post-print – author version

  This was defined to be the definitive version/form of the Work, after peer review, which has been accepted for publication, for which copyright has been assigned or a licence agreement has been signed. This version is produced by the author, with all peer-review comments and revisions integrated into the text.
- Type 3 Post-print – publisher version

  This was defined to be the definitive version/form of the Work, after peer review, which has been accepted for publication, for which copyright has been assigned or a licence agreement has been signed and with the publisher's copy-editing, formatting and production in place, i.e., it may be the publisher's PDF.

Each type is assigned a self-archiving status, comprising of one of the following:

- Status 1 – Yes

  This version of the Work can be self-archived.
- Status 2 **–** No

  This version of the Work cannot be self-archived.

- Status 3 **–** Unclear

It is not clear whether this version of the Work can be self-archived

Differentiating the work of authors into pre-print and post-print has now become standard practice in the scholarly communication field. As a result of their wide-spread use, it was agreed not to change this terminology. However, problems can arise due to different interpretations of pre-print and post-print. An example of this is the differing views of publishers and authors on what constitutes a post-print. Whereas some publishers see post-prints as the work after the *publishing* process, some authors see post-prints as the work after the *peer-review* process. The existing SHERPA/RoMEO database uses the latter definition, and this will continue in the CKB as the resource is designed mainly for use by academics rather than publishers.

Using the basic distinction of pre-print and post-print as the first entry point simplifies the description of a publisher's self-archiving policy, as it is the use of each of these which needs to be described. It also reflects the first question that authors ask when it comes to self-archiving, which is "What am I allowed to self-archive - pre-prints, post-prints, or, indeed, nothing at all?" It is then that they will want to find out details such as where the work can be archived, when it can be archived, and if there are any other conditions and restrictions in place.

There are many stages to a pre-print, including work before submission to the publisher, work submitted to and under consideration by the publisher, and work accepted for publication by the publisher, subject to completing the peer-review process, though a publishing agreement has not yet been signed. However, it was decided that in order to simplify the controlled vocabulary these different stages of a pre-print would be treated as one category, resulting in just the one pre-print type, as defined above.

It should be noted that a post-print has been accepted for publication, has had copyright for it assigned or a licence agreement signed, and has all peer-review comments and revisions integrated into the text (as defined above). However it does not necessarily mean that it has been published. A distinction must also be made between the author's version of the post-print and the publisher's version, as these are two different works, which may have different conditions and restrictions attached. In many cases, publishers do not allow their PDF version of the article, i.e., with the publisher's copy-editing and formatting in place, to be self-archived. This distinction therefore needs to be represented prominently in the database. It is interesting to note that many academic authors see the publisher's PDF version as the most convenient version of the article and in many cases as the definitive version. Many authors do not keep their own version of their post-print for self-archiving. By pointing out this distinction, the CKB should help ensure academics are aware of the need for authors to retain their final version.

Another interesting point to note is that many publishers do not specifically state which version of the post-print may be self-archived. In these cases, the publishers had to be contacted so as to clarify the situation. Where this clarification was not obtained the status of the post-print – publisher version was recorded on the database as being 'unclear'. It would be useful if, in future, publishers could make the distinction clear in their self-archiving policies and CTAs.

5.2. Conditions of Self-Archiving

The 'conditions' category consists of requirements which publishers insist are met in order to self-archive, but which do not prevent an author from self-archiving now.

The first condition that many CTAs stipulate is where work may be self-archived.

5.2.1 Where can work be self-archived?

The process of choosing which 'where' terms to use was particularly complex as publishers provide many different variations of 'where' terms, but hardly any give actual definitions. Terms used in CTAs included 'homepage', 'author's webpage', 'institutional webpages', 'institutional repositories' and many others. Therefore, it was necessary to determine what publishers actually meant when using these terms by attempting to logically interpret their intentions.

The result of this consideration was that 'where' terms represented in the controlled vocabulary are simply:
c_a. Web site
c_b. Non-commercial Server
This means that the server must not be for commercial use and does not depend on payment for access, subscription and membership fees.
c_c. 'What & Where' Free-Text Field
This field contains more detailed information on the above 'what' and 'where' conditions, e.g., the actual terms used in the CTAs.
(note: The 'c_' and 'r_' codes, used for 'conditions' and 'restrictions' respectively, are the machine readable codes that are used in the XML version of the controlled vocabulary.)

The decision to use just the terms 'Web site' and 'non-commercial sever' was based on the following considerations:

5.2.2 Is there a distinction between a Web site, Web page or Homepage?

It was felt that a web-site implies a collection of web pages mounted on a web server. Mounting a Work on a web site effectively means placing the Work (e.g. a PDF of the article) on the web server and linking to it from one or more web pages. Mounting a work on a web page effectively means the same thing.
Certain publishers have specified that a Work must be mounted on an author or institutional 'homepage'. A 'homepage' is commonly taken to mean the 'root' of a web site or part of a web site. Thus an author's homepage may not necessarily be the root of the website but an area within the site that the author manages; or it may be dedicated to information directly relating to the author. In the latter case, the author may or may not have direct editorial control. In an institutional context an individual academic may have their own 'homepage' which sits within a larger structure of pages describing an entire Department or University. In addition some publishers qualify the type of homepage by indicating 'personal', 'departmental' or 'professional'. However there are inconsistencies with this approach in that an individual's web pages which are mounted on an institutional web site often contain both personal and professional information and may be managed by either the author, the institution or, in many instances, both. Similarly many authors maintain their own web pages on non-institutional servers, which may be 'commercial' servers, that contain a mixture of personal and/or professional information. It is therefore difficult to define any clear distinctions for many of these terms. Unfortunately publishers who use these terms do not provide clear definitions for them, therefore it was felt that the controlled vocabulary should not provide a distinction between personal and professional web pages. However it was felt that if a publisher specifies that a Work may be mounted on a homepage, then what is meant is that the work be mounted on a web server and linked to from a 'homepage' within the site. The reason why publishers specify 'homepage' rather than website is unclear – it may be that they wish the work to be linked to from a prominent page, rather than buried deep within the website.
Some publishers also specify that a work may be mounted on an 'author's web site'. Again the meaning of the term is not clear. An author's web site may be regarded as one which is directly under the author's control: however, an 'author's web site' may also be deemed to consist of a homepage and sub-pages mounted on a departmentally controlled server. The distinction between 'author's-website', 'website', 'homepage' etc is therefore unclear.
However in practical terms, given the distributed nature of the web, it matters little where the actual work is located. The Work could be located on a web server owned and run by an individual author, however a link to that work could be made available from the author's institutional pages. In these circumstances it is clear that the work is 'mounted' on a personal web server , however it is accessible through an institutional webpage. To take this argument further it is perfectly possible for an Institutional Repository to contain the title, author and metadata describing a Work without hosting the Work itself. If the metadata contained a link to the Work then the Institutional Repository can be searched and the Work returned as a 'hit', yet the Work is accessed via a link from the Institutional Repository to the

actual web server where the work is physically mounted. In this case the work is mounted on a personal web server but in terms of the practicalities of accessing the Work it is available through an Institutional Repository.
Once the work is mounted on an accessible server, then links to it can be created by anyone and placed on any page: that is the nature of the web. Restricting where a link is displayed (for example, only to be linked from an author's homepage), is to misunderstand the way that the web works. While the author or institution may feel constrained by such a condition, it would be impossible to police compliance by others. Something is either available through the web and all that that entails - or it is not.
It was due to reasons such as these that it was decided to make no distinction between the various terms of website, webpage, Institutional Repository etc. To reinforce this argument it should be noted that an Institutional Repository is, in fact, simply a type of website, therefore the controlled vocabulary makes no distinction between these terms. However it should be noted that although no different categorisation is given between the terms in the controlled vocabulary, the CKB does report the actual terms used by the publisher so that users of the CKB can make their own judgements if necessary.
One exception to this decision was to make a specific category for 'non-commercial servers'. Many publishers do specify that a Work can only be mounted on a 'non-commercial server' and it was felt that this was a reasonable distinction to make, as this prevents Works from being mounted on rival publisher's websites or servers that require payment for access.

5.2.3 Should secure networks be included?

It was decided that mounting work somewhere which requires an authentication mechanism was not deemed to be Open Access (OA) self-archiving. This is because networks such as intranets and electronic reserves or servers where access is restricted by password are a way of controlling access to specific groups of people. Any restrictive authentication does not constitute OA self-archiving. Because of this no controlled vocabulary term was created for these networks, however it was still felt worthwhile to report this information in the database in the 'additional information' free-text field, thereby allowing users to fully understand the publisher's policy.

5.2.4 Does being told where to specifically mount your work constitute Open Access?

Some publishers specify where an author may mount their work, whether it be a digital archive or repository, or a named online version of a particular journal, in which authors often have to pay in order to make their work openly accessible. Originally, the controlled vocabulary contained the terms 'specified digital archive', (e.g., PubMed Central [4], arXiv [5]), and 'specified online version of the journal' (e.g., Cambridge Journals Online [6]). Both of these terms then had an 'exclusive' and 'non-exclusive' option, depending on whether the publisher restricted self-archiving to the specified place only, or whether they specified a place, but also allowed self-archiving elsewhere.
However, the exclusive option does not constitute true Open Access self-archiving, as authors do not have the right to do what they want with their Work, for example to mount it on an IR. As the CKB is a resource which promotes Open Access self-archiving where authors are free to choose where to mount work, it was felt that all details on specified archives/online journals would appear in the 'additional information' free-text field, but that if a publisher stipulated this restriction then the publisher would not be considered a "green" publisher, i.e., one that supports OA. (*More details about the logic behind the colour-coding of publishers is available[12])*
However, many publishers specify that the work should not be mounted on a commercial server. This leads into consideration of whether open access should include the right for others to be able to commercially exploit work, which is a continuing debate. For the purpose of increasing access, such a condition is not unduly restrictive. As many publishers make a point of requiring this, it was felt that this should be represented in the controlled vocabulary, and that if this was the only restriction, then the publisher could still be considered to be a "green" publisher.

5.2.5 Other Conditions for Self Archiving

There are various other conditions which publishers ask to be met for an author to self-archive. These conditions were included in the controlled vocabulary and are:

Electronic Links

c_d1. To the publisher's version of the Work on their web site.
c_d2. To the publisher's online abstract of the Work.
c_d3. To the journal's home page/web site.
c_d4. To the DOI (Digital Object Identifier) of the Work.

It was decided to keep 'c_d1' separate from 'c_d4' as a DOI is not guaranteed to link solely to the publisher's version of the Work. Also, it is possible that the Work on the publisher's Web site may not have a DOI.

Copyright Acknowledgement

c_e1. The copyright holder of the Work must be acknowledged.
c_e2. A proper/specified copyright notice must be given.

There is a distinction made between just stating someone is the copyright holder of the Work and requiring a specified copyright notice. If a publisher provides the wording for a copyright notice, this is placed in the 'specific statement' free-text field (see below).Where available, a link to the URL giving the specified copyright notice is also supplied.

First Publication Credit

c_f1. Acknowledge the original, i.e. published, source of the published Work.
c_f2. Give a full citation/bibliographic reference to the published Work

Some publishers used the term 'give a short bibliographic reference', however it was decided to incorporate this into the 'full bibliographic citation' category wherever any form of reference was specified as this will improve the clarity and usability of archives. However, there is still a distinction to be made between giving a citation and just acknowledging the original source of the published Work, especially as in many cases publishers do not specify exactly how they want to be acknowledged.

Label the Stage in the Publishing Process

c_g1. The archived Work must be labelled as being under review by or submitted to a [named] journal.
c_g2. The archived Work must be labelled to indicate it has been accepted by a [named] journal.
c_g3. The archived Work must be labelled as 'In press' or to be published by a [named] journal.

Some CTAs required 'labelling the Work as having been published in a [named] journal'. This was interpreted as logically equivalent to 'c_f1' (see above). It was decided to keep 'c_f1' rather than have it in the 'c_g' section as 'acknowledging the source' and 'giving a full citation' both refer to making a reference to the official publication of the Work. Also,'c_g1' to 'c_g3' are explicit enough to cover the stages of the publication process before actual publication. Some publishers used the term 'label the Work having not yet been published in a [named] journal', however this was not used, as it was taken to mean the same as 'c_g3'.

Version

c_h1. It must be stated that the archived Work may not exactly replicate the published Work found in the journal.
c_h2. It must be stated that the archived Work has been published in a revised form.

These two terms often have specified statements associated with them. These appear in the 'specific statement' free-text field (see below). It seems likely that publishers require these conditions so as to make people want to read what is currently considered to be the published version in its journal, therefore this does not alter the publisher's self-archiving status.

Conditions Of Reuse

c_i1. Unrestricted reuse acknowledging source and copyright

c_i2. Unrestricted reuse for non-commercial purposes
c_i3. Unrestricted reuse for non-commercial personal use.
c_i4. Unrestricted reuse for non-commercial educational purposes
c_i5. Unrestricted reuse for non-commercial research communication purposes
c_i6. Unrestricted reuse for redistribution
c_i7. Unrestricted reuse for non-commercial redistribution
c_i8. Unrestricted reuse for purposes of data mining
c_i9. Unrestricted reuse for republication.
c_i10. The archived Work must not be used for any systematic external distribution by a third party (e.g. listserve or database connected to a public access server).
Some publisher's used the phrase 'unrestricted reuse', however this was not included in the controlled vocabulary as it is neither a condition or a restriction because it does not restrict the author from self-archiving. It was therefore felt to be redundant. It could be argued that a CTA that explicitly indicates that 'unrestricted reuse' is acceptable is different from a CTA that is silent on this matter. In future the CKB may consider maintaining this distinction when analysing CTAs. 'c_i10' was originally considered as a separate category in the controlled vocabulary, but was added to this section as it is considered a condition of reuse. Originally, the conditions of reuse were structured into intersecting sets, i.e., with classes and sub-classes, but this made the vocabulary inconsistent with the CKB's simple two-tier structure. Therefore, the conditions were re-structured, starting with the least restrictive ('unrestricted reuse acknowledging source and copyright') and numbered sequentially from 'c_i1.' to 'c_i10'.

Reuse Statement
c_j1. The archived Work must state that further re-use is permitted.
c_j2. The archived Work must state that further re-use is <u>not</u> permitted.
c_j3. The archived Work must state that the permission of the copyright holder must be obtained if the Work is to be reused.
These terms are often accompanied with a specific statement.(See 'Specified Statements' below.)

Miscellaneous Conditions
c_k. The author must notify the publisher if an archived Work is to be updated or replaced with the published post-print.
c_l. This version of the Work must <u>not</u> be replaced or updated to make it identical in content to the final published post-print.
The focus of 'c_l.' is on *content*. Therefore if a publisher specifies this condition, it means that only the pre-print can be self-archived, however it does not mean that a separate addendum to the pre-print, potentially acting on the publisher's peer-review comments and revisions, cannot be posted.

Specified Statements
It was decided to put the specified statement options at the end of the 'conditions' section for practicality, with all the individual conditions having being covered first in the section.
c_m. The publisher requires and provides a specified statement/set phrase to be added to the archived Work.
c_n. 'Specific Statement' Free-Text Field
This field contains the actual statement/set phrase as specified in the CTA, and/or an electronic link to it.
In many cases, specified statements are provided by publishers which must be included on the work. These may correspond to any of the conditions given in the controlled vocabulary. An electronic link to the statement could also be provided. This not only saves time for authors or repository administrators but also encourages academics and IR administrators to consult the original CTA.

The decision was made to have a term stating a specified statement was required ('c_m'), as well as having a 'specific statement' free-text field ('c_n').

‘Conditions’ Free-Text Field
c_o. ‘Conditions’ Free-Text Field
This free-text field can contain further details on the conditions given for self-archiving, which may include such details as where a specified statement must appear and the publisher’s URL that the self-archived work must link to.

5.2.6 Future Growth

Some of the condition terms which only apply to one or two publishers, such as ‘the author must provide the publisher with the electronic address of the primary electronic posting’ were placed in the relevant publisher’s ‘conditions’ free-text field so as not to add to the number of entries in the controlled vocabulary. If more publishers specifically request such actions, then these may be added to the ‘conditions’ terms in the controlled vocabulary. It is, therefore, important to monitor what is added to the free-text fields.

5.3. Restrictions on Self-Archiving

Restrictions are terms which prevent an author from self-archiving immediately or which have an effect on long term archiving. Restrictions are more prohibitive than conditions. An example of a restriction is ‘the pre-print must be removed on submission to the publisher’ (‘r_b1’). This particular restriction prevents the pre-print from being self-archived indefinitely. The controlled vocabulary for restrictions was therefore developed as follows.

5.3.1 Formal Permission

r_a Formal Permission
Formal permission from the publisher must be sought if the Work is to be posted electronically. The resulting permission may involve a fee to be paid to the publisher to copy or transmit the article. Originally the vocabulary made a distinction between formal permission involving a fee to be paid to the publisher and formal permission without a fee. However it was felt that although it may be interesting to retain this information, for the purposes of self archiving, it was *not* necessary to distinguish between the two. Therefore it was decided just to state that formal permission may involve a fee.

5.3.2 Work Removal

r_b1. The pre-print must be removed on submission to the publisher.
r_b2. The pre-print must be removed until it has been accepted or rejected by the publisher.
r_b3. The pre-print must be removed when it has been accepted by the publisher.
r_b4. The pre-print must be removed when the copyright has been transferred to the publisher.
r_b5. The previous version of the Work must be removed on publication of the post-print.
Referring to r_b5: it was decided that ‘previous version’ does not necessarily refer to the pre-print. It may refer to the author’s version of the post-print, before the publisher version has actually been published in the journal.

5.3.3 Work Replacement

r_c1. The previous version of the Work must be replaced with its abstract and full citation.
r_c2. The previous version of the Work must be replaced with a link to the toll-free published article on the publisher’s Web site.
Here again ‘previous version’ rather than pre-print is used as this could refer to the author’s version of the post-print after its acceptance but before its publication in the publisher’s PDF version.

5.3.4 'When' Restrictions

r_d- The Work may be self-archived before its publication in the journal.
r_d0. The Work may be self-archived only on its publication in the journal.
r_d1. The Work may be self-archived 1 month after publication in the journal.
…
r_d12. The Work may be self-archived 12 months after publication in the journal.
r_d13. The Work may be self-archived at another specified time. (Please specify in r_e.)

Whilst 'where' a Work can be self-archived is judged to be a condition, 'when' it can be self-archived is judged to be a potential restriction, as the publisher may prevent the author from self-archiving through stipulating an embargo after it has been published in a journal. In many cases, this time delay is 6 months or one year. We have become aware of an increasing trend towards publishers imposing such embargoes. This is no doubt because they are concerned about a perceived threat posed to their current business model by IRs. It may also be influenced by the policies of research funders such as the US National Institutes of Heath (NIH) [7] or the Wellcome Trust [8], indicating that such embargoes are acceptable to them. Embargoes mean that the article is ONLY available by subscription for a certain period of time. However, embargoes are not compatible with the spirit and meaning of OA.

It was decided to represent embargoes from one month to a year, with an option for over a year. In this instance, the exact duration of the embargo would need to be recorded in the 'restrictions' free-text field. Providing a range of times in the controlled vocabulary allows for possible future changes in publishers' CTAs. It is, of course, important to ensure that the CKB can easily be extended to cover future terms and definitions developed by publishers.

As the embargoes are to be found in the restrictions section, a decision was made to place all the 'when' options here for consistency. Therefore r_d-, self-archiving 'before publication in the journal' is also found in the 'when' restriction section, even though this is not a restriction and is, in fact, an indication that publishers have considered the importance that immediate self archiving can have on the impact of journal articles.

One interesting point to note is that the 'when' option of 'self-archiving only on publication' ('r_d0') is considered as a restriction. This is because this restriction means that the author has to wait until their Work is published before they can self-archive it. Although this option is not a severe restriction, it still should be theoretically regarded as a restriction as immediate access to peer refereed research may be delayed by the publication process.

5.3.5 'Restrictions' Free-Text Field

r_e 'Restrictions' Free-Text Field

This can contain further details on the restrictions given for self-archiving, which may include such information as a specified embargo period that is of more than 12 months.

5.4. 'Additional Information' Free-Text Field

This can contain any further information which it may be important to know, but which does not apply specifically to any of the free text field categories above. This information may be of peripheral interest to self-archiving, such as only allowing work to be made available through a secure network, or in a specific digital archive.

In the majority of cases the additional information applies to all three 'types' represented in the database, i.e., 'pre-print', 'post-print – author version' and 'post-print – publisher version', rather than just to one specific 'type'.

In practice, the 'additional information' free text field has been mainly used to record the following information:

- The publisher offers an Open Access option on payment of an additional fee;
- The publisher allows self-archiving under the Open Access policy of the US NIH.

Examples of additional information include:
“Authors of accepted peer-reviewed articles may choose to pay a fee in order for their published article to be made freely accessible to all via our online journals platform, Blackwell Synergy … the Online Open fee will be fixed at US$2600, 1950 Euros or £1300 (plus VAT where applicable). Any additional standard publication charges will also apply, such as for color images or supplementary datasets.” [9]
“An author or group of authors may post without further permission, on their own personal or organizational Web site(s) the title, authors, and full abstract of their paper(s), providing the posting cites the GSA publication in which the material appears and the citation includes the address line: "Geological Society of America, P.O. Box 9140, Boulder, CO 80301-9140 USA (http://www.geosociety.org)," and also providing the abstract as posted is identical to that which appears in the GSA publication.” [10]

6. Conclusion

The aim of devising the controlled vocabulary was to enable journal publishers’ CTAs to be systematically analysed. Self-archiving terms were identified, defined, and structured to form the controlled vocabulary. This process has led to a simplification in understanding the terminology used by publishers, and it facilitates a user-friendly approach to representing the terms in the CKB. It is not clear whether it will lead to a simplification of the actual terminology that publishers use, but it would be helpful if it did, as this would clarify the situation for authors and repository administrators.
Although a definitive controlled vocabulary has been developed, it is important to keep it up to date in the light of changes to publisher practices. The scholarly communication environment is continually changing, with new publishing models, including the ‘author-pays’ model, and new Open Access initiatives and policies, such as that of the US NIH, contributing to a shift in research communication practices. Such developments need to be taken into account when maintaining the controlled vocabulary.
As a result of the possible changes to publishing policies, the controlled vocabulary must be able to be easily expanded to include new terms, and maybe even new categories. A certain amount of forethought has already been employed in the development of the controlled vocabulary. This is illustrated in particular by the ‘when’ terms, with a number of embargo periods, from one month to over one year, being available.
The controlled vocabulary has been used to analyse all the SHERPA/RoMEO publishers; these represent the majority of the larger scholarly publishers world-wide. However, as more publishers’ CTAs get analysed, the choice and definitions of terms may need to be reassessed. This has already been indicated with regards to the ‘what’ terms, which may already not be as extensive as they could be. Here, the research being carried out by the JISC VERSIONS project [11] on the lifecycle of an academic research paper will be of major interest. Nonetheless, the controlled vocabulary described in this paper, will act as the basis for analysing publisher CTAs and for classifying publishers’ self-archiving policies as “green”, “blue”, “yellow” or “white”.
The existing SHERPA/RoMEO database is in the process of being upgraded so that the underlying database fully supports a detailed CTA analysis using the controlled vocabulary. This process will occur throughout 2007. As of June 2007, three hundred and three publishers’ CTAs have been analysed by SHERPA/RoMEO. Table 1 shows the statistics outlining the number of publishers falling into each colour category.

Table 1: The number of publisher's in each colour category

RoMEO colour	Archiving policy	Publishers	%
green	**can archive pre-print and post-print**	**110**	**36**
blue	**can archive post-print (ie final draft post-refereeing)**	**74**	**24**
yellow	**can archive pre-print (ie pre-refereeing)**	**31**	**10**
white	**archiving not formally supported**	**88**	**29**

*Note: 73% of publishers allow some form of self archiving

In addtition, The controlled vocabulary will be important in a number of areas. Although a central searchable database of CTAs (as provided by the existing SHERPA/RoMEO database) seems the most appropriate architecture, the controlled vocabulary will enable consistency of CTA analysis and may enable the analysis of CTAs to be distributed/decentralised. It is only if a standardised vocabulary exists that CTA analysis can be performed objectively and with rigour. The controlled vocabulary may enable multiple partners to perform the analysis and, to this end, it should also facilitate the integration of foreign language CTAs into the SHERPA/RoMAO database.
In future, if considered appropriate, the controlled vocabulary could enable CTAs to be represented and disseminated in a machine understandable format. It would be possible for the CKB to provide an API that would accept a web-based request for information and return the terms and conditions of a CTA in XML form. This would be a first step to enabling a repository of machine readable CTAs (or at least a repository of machine readable metadata descriptions of CTAs). This would enable repositories to provide automated services to authors or depositors of e-prints informing them of the requirements of a particular CTA (e.g. that a specific statement must be appended to a postprint-author version before submission to the repository).

7. Acknowledgements

We wish to thank JISC and SURF for funding this research, and Mike Gardner for much technical assistance.

8. References

[1] JISC-SURF 'Partnering on Copyright' programme (2006). Available at: http://www.jisc.ac.uk/whatwedo/projects/programme_jiscsurfipr.aspx (accessed 17 Jnue 2007).

[2]. RoMEO (Rights MEtadata for Open Archiving) Project (2003). Available at: http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html (accessed 15 June 2007).

[3] SHERPA/RoMEO Database of Publishers' Self-Archiving Policies. Available at: http://www.sherpa.ac.uk/romeo.php (accessed 20 June 2007).

[4] PubMed Central (2005). Available at: http://www.pubmedcentral.nih.gov/ (accessed 16 February 2007).

[5] arXiv Available at: http://arxiv.org/ (accessed 16 June 2007).

[6] Cambridge Journals Online (2006). Available at: http://journals.cambridge.org/action/home (accessed 20 June2007).

[7] US National Institutes of Health (NIH) Public Access Policy (2005). Available at: http://publicaccess.nih.gov/policy.htm (accessed 16 June 2007).

[8] Wellcome Trust Position Statement in Support of Open and Unrestricted Access to Published Research (2006). Available at: http://www.wellcome.ac.uk/doc_WTD002766.html (accessed 10 June 2007).

[9] Blackwell Publishing, *About Online Open*. Available at: http://www.blackwellpublishing.com/static/onlineopen.asp (accessed 10 June 2007).

[10] Geological Society of America, Copyright Information(2005). Available at: http://www.geosociety.org/pubs/copyrt.htm#web (accessed 14 June 2007).

[11] JISC VERSIONS (Versions of Eprints – user Requirements Study and Investigation Of the Need for Standards) Project (2005). Available at: http://www.lse.ac.uk/library/versions/ (accessed 14 June 2007).

[12] C. Jenkins, S. Probets, C. Oppenheim and B. Hubbard, 'RoMEO Studies 8: Self-archiving. The logic behind the colour coding used in the Copyright Knowledge Bank', Program – electronic library and information systems. 41(2), 2007 pp124-133.

Appendix : The Copyright Knowledge Bank - Controlled Vocabulary and Definitions

Publisher's Details
ID
Name
Home URL
CTA/Self-archiving policy details URL

Publisher's Copyright/Self-archiving Status
A1 – The publisher has the same self-archiving policy for all its journals.
A2 – The publisher has different self-archiving policies for different journals.

B1 – The publisher requires the author to transfer their copyright.
B2 – The publisher does not require the author to transfer their copyright.
B3 – The author is explicitly given the option to retain their copyright.

C1 – Self archiving is not formally supported. [WHITE]
C2 – Pre-prints (i.e., during peer-review) can be self-archived. [YELLOW]
C3 – Post-prints (i.e., after peer-review) can be self-archived. [BLUE]
C4 – Both pre-prints and post-prints can be self-archived. [GREEN]

What Can Be Archived?
Type
1- Pre-print
The primary, draft version/form of the Work, up to and during peer-review.

2- Post-print – author version
The definitive version/form of the Work, after peer-review, which has been accepted for publication, for which copyright has been assigned or a license agreement has been signed. This version is produced by the author, with all peer-review comments and revisions integrated into the text.

3 – Post-print – publisher version
The definitive version/form of the Work, after peer-review, which has been accepted for publication, for which copyright has been assigned or a license agreement has been signed. This version has the publisher's copy-editing, formatting and production in place, i.e., is in publisher's PDF form.

Status
1 – Yes
This version of the Work can be self-archived.

2 – No
This version of the Work cannot be self-archived.

3 – Unclear
It is not clear whether this version of the Work can be self-archived.

Conditions - Do not prevent an author archiving now
c_a Web site

c_b. Noncommercial Server

The server must not be for commercial use and does not depend on payment for access, subscription and membership fees.

c_c. 'What & Where' Free-Text Field
Contains more detailed information on the above 'where' conditions, e.g. actual terms used..

Electronic Links
c_d1. To the publisher's version of the Work on their website.
c_d2. To the publisher's online abstract of the Work.
c_d3. To the journal's home page/website.
c_d4. To the DOI (Digital Object Identifier) of the Work.

Copyright Acknowledgement
c_e1. The copyright holder of the Work must be acknowledged.
c_e2. A proper/specified copyright notice must be given.

First publication credit
c_f1. Acknowledge the original, i.e. published, source of the published Work.
c_f2. Give a full citation/bibliographic reference to the published Work

Label The Stage In The Publishing Process
c_g1. Under review by or submitted to a [named] journal.
c_g2. Been accepted by a [named] journal.
c_g3. 'In press' or to be published by a [named] journal.

Definitive Version
c_h1. It must be stated that the archived Work may not exactly replicate the published Work found in the journal.
c_h2. It must be stated that the archived Work has been published in a revised form.

Conditions Of Reuse
c_i1. Unrestricted reuse acknowledging source and copyright
c_i2. Unrestricted reuse for non-commercial purposes
c_i3. Unrestricted reuse for non-commercial personal use.
c_i4. Unrestricted reuse for non-commercial educational purposes
c_i5. Unrestricted reuse for non-commercial research communication purposes
c_i6. Unrestricted reuse for redistribution
c_i7. Unrestricted reuse for non-commercial redistribution
c_i8. Unrestricted reuse for purposes of data mining
c_i9. Unrestricted reuse for republication.
c_i10. The archived Work must not be used for any systematic external distribution by a third party (e.g. listserve or database connected to a public access server).

Reuse Statement
c_j1. The archived Work must state that further re-use is permitted.
c_j2. The archived Work must state that further re-use is not permitted.
c_j3. The archived Work must state that the permission of the copyright holder must be obtained if the Work is to be reused.

c_k. The author must notify the publisher if an archived Work is to be updated or replaced with the published post-print.

c_l This version of the Work must not be replaced or updated to make it identical in content to the final published post-print.

c_m. The publisher requires and provides a specified statement/set phrase to be added to the archived Work.

c_n. The actual statement/set phrase as specified in the CTA, and/or a link to this.

c_o. 'Conditions' Free-Text Field
Contains further information on the above conditions, e.g. where to locate specific statements, and more details on conditions of use.

Restrictions – Prevent an author from archiving (in the long-term)
Formal Permission
r_a. Formal permission from the publisher must be sought if the Work is to be posted electronically. The resulting permission may involve a fee to be paid to the publisher to copy or transmit the article.

Work Removal
r_b1. The pre-print must be removed on submission to the publisher.
r_b2. The pre-print must be removed until it has been accepted or rejected by the publisher.
r_b3. The pre-print must be removed when it has been accepted by the publisher.
r_b4. The pre-print must be removed when the copyright has been transferred to the publisher.
r_b5. The previous version of the Work must be removed on publication of the post-print.

Work Replacement
r_c1. The previous version of the Work must be replaced with its abstract and full citation.
r_c2. The previous version of the Work must be replaced with a link to the toll-free published article on the publisher's Web site.

When (Embargoes)
r_d- The Work may be self-archived before its publication in the journal.
r_d0. The Work can only be self-archived on its publication in the journal.
r_d1. The Work can only be self-archived 1 month after publication in the journal.
r_d2. The Work can only be self-archived 2 months after publication in the journal.
r_d3. The Work can only be self-archived 3 months after publication in the journal.
r_d4. The Work can only be self-archived 4 months after publication in the journal.
r_d5. The Work can only be self-archived 5 months after publication in the journal.
r_d6. The Work can only be self-archived 6 months after publication in the journal.
r_d7. The Work can only be self-archived 7 months after publication in the journal.
r_d8. The Work can only be self-archived 8 months after publication in the journal.
r_d9. The Work can only be self-archived 9 months after publication in the journal.
r_d10. The Work can only be self-archived 10 months after publication in the journal.
r_d11. The Work can only be self-archived 11 months after publication in the journal.
r_d12. The Work can only be self-archived 12 months after publication in the journal.
r_d13. The Work can only be self-archived at another specified time. (Please specify in r_e.)

'Restrictions' Free-Text Field
r_e. Contains further information on the above restrictions.

Additional Information (Free-Text Field)
Additional information and interesting points which do not fit into any of the above fields.